Available online at www.sciencedirect.com

ScienceDirect

Theriogenology

Theriogenology 75 (2011) 783–795

www.theriojournal.com

Invited review

Statistical Series: Opportunities and challenges of sperm motility subpopulation analysis[1]

Felipe Martínez-Pastor[a,b,*], E. Jorge Tizado[c], J. Julian Garde[d], Luis Anel[a,e], Paulino de Paz[a,b]

[a] *INDEGSAL, University of León, 24071 León, Spain*

[b] *Molecular Biology (Cell Biology), University of León, 24071, León, Spain*

[c] *Zoology, Biodiversity and Environmental Management, University of León, 24071 León, Spain*

[d] *Biology of Reproduction Group, National Wildlife Research Institute (IREC), CSIC-UCLM-JCCM, and Institute for Regional Development (IDR), 02071, Albacete, Spain, 02071, Albacete, Spain*

[e] *Animal Reproduction and Obstetrics, University of León, 24071, León, Spain*

Received 20 September 2010; received in revised form 5 November 2010; accepted 17 November 2010

Abstract

Computer-assisted sperm analysis (CASA) allows assessing the motility of individual spermatozoa, generating huge datasets. These datasets can be analyzed using data mining techniques such as cluster analysis, to group the spermatozoa in subpopulations with biological meaning. This review considers the use of statistical techniques for clustering CASA data, their challenges and possibilities. There are many clustering approaches potentially useful for grouping sperm motility data, but some options may be more appropriate than others. Future development should focus not only in improvements of subpopulation analysis, but also in finding consistent biological meanings for these subpopulations.

Keywords: CASA; Automated semen analysis; Sperm subpopulations; Multivariate analysis; Cluster analysis

Contents

* Corresponding author. Tel.: +34 987 291 000; fax: +34 987 291 000.
E-mail address: fmarp@unileon.es (F. Martínez-Pastor).

[1] This article is part of the Statistical Series guest-edited by Szabolcs Nagy.

0093-691X © 2011 Elsevier Inc. Open access under CC BY-NC-ND license.
doi:10.1016/j.theriogenology.2010.11.034

1. Introduction

Many studies have showed that the spermatozoon is a dynamic cell, with active biochemical pathways that modify the sperm physiology throughout maturation, ejaculation, transport in the female genital tract and fertilization. Flagellar beating is affected by these changes [1–3], thus spermatozoa show different swimming patterns in the epididymis, seminal plasma, cervical mucus, oviduct (capacitation) and while penetrating the oocyte vestments [4]. In many ways, motility integrates the biochemical events occurring in the spermatozoa. Moreover, sperm samples are heterogeneous, implying that spermatozoa with different motility coexist in the same ejaculate [5–8]. Therefore, the analysis of sperm subpopulations based on motility characteristics may help to assess the status of the sperm sample and its fertility potential, exploiting the heterogeneity of sperm samples.

This kind of detailed analysis was allowed by the spread of computer-assisted sperm assessment (CASA). The term CASA defines hardware and software designed to acquire and digitize successive images of sperm, process and analyze the image sequence, and output the information of the kinematics of individual spermatozoa. CASA details will not be treated here, but other reviews have ample information on the topic [4,9–15]. Despite of its advantages, some authors have warned about the misuse of CASA results [16]. Moreover, the capacity of CASA for generating huge datasets comprising motility data from thousands of spermatozoa has been overlooked in favor of the summary statistics provided by the software, which do not show the intrinsic variability of the sample.

Cluster analysis is a technique for statistical data analysis that allows unsupervised grouping of observations into subsets (called clusters), so that observations in the same cluster are similar depending on a given criteria [17–19]. "Unsupervised" implies that there is not an *a priori* grouped dataset to guide the grouping. Cluster analysis is also a multivariate technique, because the information for grouping observations is provided by several to many descriptors characterizing the observations. Therefore, cluster analysis is perfectly suited to resolve the heterogeneity of sperm motility data in discrete subpopulations, helping to take advantage of the information contained in CASA datasets.

In this review, we start by giving a quick outlook to previous studies on sperm motility subpopulations. However, most of this review aims at explaining how to perform and validate cluster analysis, in the context of CASA data. Nevertheless, because of the complexity of the topic, this explanation is far from being comprehensive, and we recommend the reader look up the bibliography for specific information. We hope that this review will inspire spermatologists to embrace new statistical techniques, and to apply them in their studies.

2. Sperm motility subpopulations and their relation with sperm quality and fertility

Many studies have explored the use of cluster analysis to identify subpopulation patterns in sperm samples. Although this review is aimed at analysis of motility data, we must keep in mind that other kinds of spermatozoa data can be used for such analyses (e.g., morphology data [20–22]). Before 2000, few studies reported using these techniques for processing CASA data [23–27]. However, some 30 articles have dealt with this topic in the last decade (Tables 1 and 2), concerning around 11 species and using different statistical approaches.

Once identified, it is easy to characterize each subpopulation accordingly to its average kinematic variables. For instance, a subpopulation with high velocity variables and high linearity variables (see Table 3) could be defined as "fast, linear", whereas another could be defined as "slow, non-linear" if these variables are low. Then, the frequencies of these subpopulations are used, rather than the kinematic variables themselves. Variations in these frequencies have been associated to individual variations among ejaculates and males [28], to sperm freezability [29], or to sperm fertility [30].

Table 1
Review of clustering methods used by different authors on CASA data (continued in Table 2). The table features the species studied, references, the number of variables entered in the analysis (displaying the variable reduction method, if any), the clustering method, and the resulting number of clusters. In some cases, specific details about the clustering algorithms are not available or unclear. Detailed information regarding proprietary software may not be available.

Species	References	Variables	Clustering method	Clusters
Bull	[57,61]	8	k-means, then hierarchical (Ward linkage)	4
Dog	[28]	2 (PCA)	k-means	11
	[68]	2 (PCA)	BIRCH, then hierarchical[a]	6
	[60]	8	k-means[b]	4
Donkey	[40,43]	6 (hierarchical clustering)	k-means[b]	4
	[58]	8	k-means[b]	4
Gazelle	[27,70]	7	partitional, then hierarchical[c]	4
Goat	[51]	3 (PCA)	k-means[b], then hierarchical (average linkage)	4
Horse	[30]	7 (hierarchical clustering)	BIRCH, then hierarchical[a]	5
	[69]	3 (subjective trimming)	BIRCH, then hierarchical[a]	4, 6
Human	[24,25]	3 (iterative)	k-means	5

[a] Two-step SPSS™ procedure. The first step is based in the BIRCH (balanced iterative reducing and hierarchical clustering) algorithm for large datasets [67]; the second step is an agglomerative hierarchical procedure (not clear which kind it uses).
[b] FASTCLUS procedure of SAS™, a fast k-means algorithm.
[c] ALOC (partitional), Flexible UPGMA (hierarchical, average linkage) and FUSE (hierarchical) modules from the PATN package.

It is interesting to note that, except for a few studies [28], most researchers have found a low number of subpopulations (3–4) when analyzing motility of sperm samples, even among quite different species. When reviewing these studies, we can find some similarities in the subpopulation patterns. In fact, the presence of a "fast and linear" subpopulation has been proposed as a good indicator of sample quality, whereas a predominant "slow and non-linear" subpopulation would be a marker of poor quality. As an example of the improvement that subpopulation analysis conveys over utilization of average CASA measurements, we can compare the outcome of two of our studies on the effects of postmortem time on epididymal spermatozoa from red deer. In a first paper [31], in which we did not use subpopulation analysis, we reported a drop of CASA variables with post-mortem time, but we missed further information from CASA data. However, when we applied cluster analysis in a second study [32], we found that the motility decrease comprised a complex dynamics involving three subpopulations, marked by the decrease of the "fast and linear" subpopulation and the increase of a "slow and non-linear" subpopulation with time. We also found that, after 48 h post-mortem, a new

Table 2
Review of clustering methods using by different authors on CASA data (continued from Table 1).

Species	Reference	Variables	Clustering method	Clusters
Marmoset	[26]	8	k-means	2
Pig	[26]	8	k-means	3
	[27,63,71]	7	partitional, then hierarchical[a]	3
	[39,42]	7 (hierarchical clustering)	k-means[b]	3
	[44]	6 (hierarchical clustering)	k-means[b]	4
	[56,58]	8	k-means[b]	4
Rabbit	[41]	7 (hierarchical clustering)	k-means[b]	4
Red deer	[29,32,50]	2 (PCA)	k-means[b], then hierarchical (average linkage)	3, 4
	[75]	4 (lower correlations)	CLARA[c], then model-based	3
Sole fish	[33]	2 (PCA)	CLARA[c]	4
	[86]	8	BIRCH, then hierarchical[d]	4

[a] ALOC (partitional), Flexible UPGMA (hierarchical, average linkage) and FUSE (hierarchical) modules from the PATN package.
[b] FASTCLUS procedure of SAS™, a fast k-means algorithm.
[c] A partitional algorithm for clustering large datasets [17].
[d] Two-step SPSS™ procedure. See footnote [a] in Table 1.

Table 3
Definitions for some standardized abbreviations of CASA kinematic parameters [4].

Acronym	Meaning
VCL	Curvilinear velocity: the time-average velocity of the sperm head along its actual trajectory.
VSL	Straight-line velocity: the time-average velocity of the sperm head along a straight line from its first detected position to its last detected position.
VAP	Average-path velocity: the time-average velocity of the sperm head along its average trajectory. The average trajectory is computed by smoothing the actual trajectory.
LIN	Linearity: the linearity of the curvilinear trajectory (VSL/VCL).
STR	Straightness: the straightness of the average path (VSL/VAP).
ALH	Amplitude of lateral head displacement: the amplitude of variations of the actual sperm-head trajectory along its average trajectory.
BCF	Beat-cross frequency: the time-average rate at which the actual sperm trajectory crosses its average path trajectory.
DNC	Dance: a measure of the pattern of sperm motion VCL×ALH).

"slow and linear" subpopulation appeared, being involved in the loss of sample quality. Taking into account the importance of the "fast and linear" subpopulation that was suggested by another study [29], the value of performing cluster analysis on these samples emerges clearly. Nevertheless, we must keep in mind that subpopulation analysis has still to spread among spermatology studies, and that conclusions about particular species and situations may not apply to others. For instance, Quintero-Moreno et al [30] showed that a "slow and linear" subpopulation seemed to be associated to the semen samples of stallions of proven fertility.

Since male-to-male differences regarding subpopulation patterns have been shown to be evident [28,33], this kind of study seems promising and bound to be developed further. In this context, there is a need to clarify the meaning of these subpopulations and to test if they correspond to a functional or physiological reality. This way, the "correct" and "wrong" subpopulations could be effectively used to assess ejaculate or male quality, and sperm work techniques could be developed so to improve the presence of the former and reduce or remove the latter. Moreover, standardization and improvement of the statistical methods used to disclose the subpopulation pattern must be carried out at the same time, in order to improve the results and facilitate comparison among studies. In the next sections we will explain the basis of these kinds of analyses, suggesting some new ways to carry them out with CASA data.

3. Preliminaries of the subpopulation analysis

3.1. Data acquisition and evaluation

A requirement for correct subpopulation studies is that the CASA system must render accurate data. It is advisable to work with high-definition cameras, capable of acquiring more than 30 frames per second (more than 50 for mammals [13] and more than 100 for fish [15]), and to use microscopes with appropriate optics and measurement chambers adapted to the species studied [9,34]. The combination of clearly acquired image-sequences and sophisticated image processing allows obtaining reliable motility parameters, resulting in richer datasets, and possibly in better clustering results.

It is important to have a good knowledge of the principles governing the operation of CASA systems, the meaning of the kinematic parameters, and the characteristics and reliability of our own CASA system [4,13]. For instance, smoothing algorithms may vary among CASA software. That means that VAP, ALH, as well as other angular or distance parameters, could be different even among versions of the same system [13] (see Table 3 for definitions of some abbreviations of CASA kinematic parameters). Moreover, as we will see later, some kinematic variables are more important than others in the clustering process, but their relative importance may vary depending on the characteristics of the CASA system and the experimental conditions. Therefore, CASA characteristics must be taken into account when identifying the most relevant kinematic variables. A combination of subjective knowledge and statistical methods can be useful to correct our original assumptions and to identify unsuspected noise in our preferred variables.

Raw datasets are usually plagued with problems for the clustering process. Typical problems comprise: lack of normality, strong skewness, outliers, data "noise", a weak clustering structure, colinearity among different variables (no independence) and differences among variables regarding internal variability. Nevertheless, we have to take into account that datasets containing clusters are expected to bear non-normal (multi-modal)

and skewed variables. Thus, the presence of such features should not be automatically taken as a sign of problematic data.

The number of motile spermatozoa included in the dataset must be large. The objective of a cluster analysis is to obtain frequencies and summary statistics from the subgroups obtained after partitioning the data, and these statistics should be reliable (that is, with narrow confidence intervals). Therefore, enough motile spermatozoa must be acquired (at least 200 per sample) in order to locate a relatively large number in each subgroup, after the clustering. A pilot experiment could be useful to estimate the results of the clustering, establishing a minimum number of motile spermatozoa that should be acquired. Nonetheless, it may be impossible to achieve that target, either because of few motile spermatozoa, to the presence of small clusters (which would force us to acquire an unrealistically large number of spermatozoa), or to experimental limitations (low sperm concentration, acquisition of few fields, etc.). If facing such problems, one must be aware of the implications of obtaining summary statistics from a low number of events, and must take that into account when interpreting the results.

Having obtained the data, the first step in the evaluation of CASA datasets is outlier pruning. Datasets should be examined for extreme or unreliable data, which could deeply affect clustering results. CASA errors are generally easy to spot and remove (e.g., records with unrealistic velocity values). Nevertheless, it is often difficult to determine if an event is a true outlier (and thus, if it should be removed) or a genuine event belonging to an underrepresented—albeit valid—cluster. An added difficulty is the multidimensionality of CASA data, requiring specific algorithms to detect outliers [35]. Interestingly, the vulnerability of typical clustering methods to outliers (e.g., k-means clustering), tending to group outliers in clusters of one or few events, can be used to remove them. A preliminary clustering step is applied to the raw dataset; then that first solution is examined, removing small clusters with extreme median values. Moreover, there are some clustering methods that can deal with noise or outliers (e.g., model-based clustering [36]).

Data transformation (sine-root, logarithmic and others [37]) may be necessary before the clustering step, especially if variables have a high skewness (detectable by descriptive statistics and histograms). Nonetheless, transformations should preserve the multimodal distribution that we would expect in CASA data. Similarly, standardization of the variables (fitting them to the same scale) is advisable. Otherwise, variables with large values (e.g., VCL, DNC) would dominate, whereas others would be underrepresented in the clustering process (e.g., ALH, BCF). The most typical standardization is the z-transformation. However, this transformation is based in the mean and standard deviation of the data, which might not be optimal for CASA variables (often non-normal). Transformations based on more robust estimates, such as the median and median absolute deviation (MAD), may be preferable.

3.2. Variable selection

CASA data are characterized by the high number of kinematic variables (from 8 to more than 20), and by the redundancy of these variables. This redundancy arises from the fact that many variables convey similar information (e.g., VCL, VAP and VSL, all of them describing spermatozoa velocity), whereas other are derived (e.g., LIN is the VSL/VCL ratio). Therefore, it is desirable to reduce the number of variables before feeding the data to the clustering algorithm, for reducing both dimensionality and redundancy. Moreover, not all variables contribute equally to defining the cluster structure, and an incorrect variable selection could result in poor clustering [38]. Selecting the appropriate variables is difficult and varies among studies (see Tables 1 and 2), due to differences among CASA systems, acquisition methodology, experimental conditions and species. Therefore, a combination of subjective knowledge and statistical methods should be applied in this step. A simple correlation analysis should disclose subsets of highly correlated variables, suggesting redundant ones. Variable clustering is another technique that can be easily applied, grouping highly related variables [30,39–46], and allowing to select a representative variable from each group for further processing. In other fields, some authors have proposed the use of iterative clustering methods to select the variable subsets that better disclose the cluster structure [47,48].

Principal component analysis (PCA) is a dimension-reducing tool that has been used for working with CASA data [28,29,32,33,49–51]. PCA replaces the variables in a multivariate data set by an uncorrelated set of derived variables (linear combinations of the initial variables) called principal components. This allows selecting only the principal components that convey most variance, thus reducing the number of variables. However, the use of principal components in cluster analysis has been criticized [52,53], because they may not capture the cluster structure. Neverthe-

less, PCA can still be a valuable tool for data exploration and variable selection [54,55].

4. Clustering and validation

4.1. Calculation of a distance matrix

The clustering process presents a high number of choices to the researcher. In fact, before carrying out the actual clustering, it is usually necessary to perform a preliminary step to generate a triangular matrix of distances between observations (in this case, among spermatozoa). These distances are a measurement of proximity among observations in the multidimensional space formed by the selected kinematic parameters: they show how similar each spermatozoon is to each other. We must point out that this step is not obvious in many statistic packages, where the calculation of the distance matrix and the clustering are presented as a single step. Some CASA studies reported used Euclidean distances [30,39,41,42,44–46,56–61], and it is likely that those not providing that information used this metrics too. It would be interesting to compare the performance of other metrics (such as Manhattan or Mahalanobis distances), because the performance of clustering methods may vary [18,19,62]. Moreover, the algorithm selected for generating the dissimilarity matrix influences the geometry of the clusters found, which may differ from the real clusters. Indeed, Euclidean distances would not be adequate in some cases [18,19].

4.2. Clustering methods

This is the core step in subpopulation analysis, and the one that most influences the number and characteristics of the groups obtained. Like other statistical methods, clustering algorithms expect that the data comply with a set of assumptions. Therefore, researchers willing to perform subpopulation analysis must take this into account when acquiring and processing CASA data, because if these assumptions are not met, results might not be correct. Likewise, a particular clustering algorithm is better suited for some kinds of dataset, while performing poorly in others. There is a huge choice of clustering methods [18], and only some of them may be useful for CASA data. For instance, CASA datasets rarely show well-separated or regular clusters, as exemplified in the Figure 1 and highlighted in several studies [25–27,29,63,64]. Therefore, methods based on the assumption of well-separated clusters

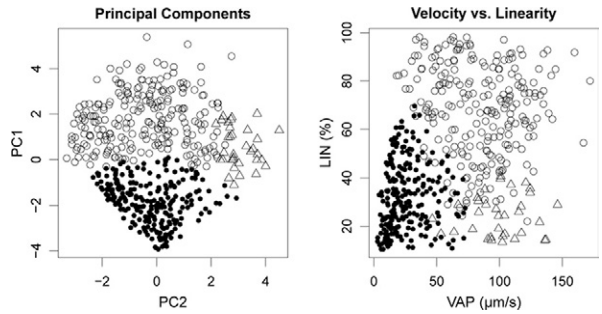

Fig. 1. Two representations of the same dataset of CASA motility data (redraw from Martinez-Pastor et al [32]), displaying the two first principal components of a PCA (left) or the VAP and LIN variables (right). Shapes identify cases belonging to three clusters in this dataset (dots, slow and non-linear; circles, fast and linear; triangles, fast and non-linear). Note the lack of regularity and the proximity of the clusters.

or center-based clusters would be less suited for that purpose [65].

Clustering methods can be grossly divided among partitional and hierarchical. However, other kinds of clustering algorithms are currently available, and could be useful for processing CASA data. Here, we will quickly review these methods, although, the reader is encouraged to consult specific references in the bibliography for algorithm details, examples, and requirements of the different methods (e.g., [17–19,53,62,65,66]).

4.2.1. Partitional (non-hierarchical) methods

In partitional methods, the final number of clusters (k) is decided by the researcher before carrying out the actual clustering. Thus, the algorithm begins assigning the observations to the k clusters, recalculating cluster membership in an iterative manner, and seeking for the optimal partitioning of the data. The k-means algorithm is the simplest and most popular partitional method, but it suffers from many drawbacks, of which researchers should be aware. The choice of initial clusters and the convergence to a global optimum can be problematic, and the algorithm is sensitive to outliers and noise. Xu and Wunsch [18] reviewed the problems and possible improvements of the k-means algorithm. There are more robust versions of the k-means method [17], including algorithms more computationally efficient, capable of processing large datasets in acceptable times. Such algorithms (e.g., CLARA and its variants [17]) are of use for CASA datasets, which often contain thousands of observations. Most CASA studies have made use of k-means methods or other partitional methods, either alone or prior to a hierachical clustering (see Tables 1 and 2).

4.2.2. Hierarchical methods

Instead of using a single-level procedure like partitional methods, hierarchical methods can be understood as a multiple-step procedure, with a single large cluster in one extreme and singleton clusters at the other. There are two subfamilies of hierarchical methods, divisive and agglomerative [65]. In divisive methods, the algorithm starts from a single cluster, and clusters are split in successive steps. Contrary, agglomerative methods start by assuming that each observation form a singleton cluster, which are joined two by two in each step. In any case, these procedures create a hierarchy of nested clusters, which can be plotted as a tree or dendrogram (very useful for descriptive purposes or for analyzing relationships among sub-clusters and choosing consistent clusters). There are quite a lot of algorithms producing hierarchical clustering, each one with its own strengths and weaknesses. Among the agglomerative algorithms, single linkage (nearest neighbor) and complete linkage (further neighbor) seem to be less suited for clustering CASA data, because these methods are sensitive to noise and expect regular clusters, respectively. Average linkage (UPGMA) or Ward's averaging method may be more appropriate. As a drawback, hierarchical methods generally require high computational resources. Therefore, in CASA studies, they have been used to further process the results of a partitional clustering (see Two-step methods below), instead of processing the raw data. Nevertheless, high-performance hierarchical clustering algorithms have been developed [18,67], and have been used in several CASA studies [68,69].

4.2.3. Two-step methods

Two clustering methods are often combined sequentially in order to get the advantages of both, especially when the second method has unpractical requirements (computation time or memory size). A fast method, with light computational requirements, can be used on the raw data to produce a relatively large number of clusters, and the centers of these clusters can be fed to the second one. The second clustering step is used to identify a set of sensible clusters, while overcoming the limitations of the first method. Generally, partitional methods are employed as the first step. The clusters produced by the partitional method are then merged in the second step by an agglomerative hierarchical method, allowing visualization of their relationships in a dendrogram [27,29,32,50,57,61,63,70,71]. The first step may also be used to identify outliers or special clusters, allowing continuation to the second step with an optimized set of clusters.

4.2.4. Other methods: fuzzy and model-based clustering

We must mention other two clustering methods with potential for being used for clustering CASA data. Fuzzy clustering [18] is characterized by not assigning absolute membership to the observations. Instead, each object may feature several degrees of membership to all clusters. This kind of classification could be useful with CASA data, because overlapping and irregular clusters are commonly present in the datasets. The other method is model-based clustering [72], which assumes that the sample comes from a mixture of several populations. With this approach, model-based clustering intends to solve some typical problems, such as detecting the number of actual clusters, choosing the most suitable clustering algorithm among a choice of those and removing outliers [36,73]. Using maximum likelihood methods, the algorithm tries to identify the best model (according to the putative cluster characteristics), and then performs an agglomerative hierarchical clustering seeking to optimize the model. Although model-based clustering performs well in small or moderately sized datasets, the computational requirements for large datasets (common when working with CASA data) are prohibitive. However, some alternative approaches have been described recently [74], allowing the use of these methods with datasets up to 100,000 cases. Working with a large database of deer spermatozoa, we used model-based clustering as the second step of a two-step clustering with good results [75].

4.3. Deciding on the number of clusters

As indicated previously, rarely there is *a priori* knowledge of how many clusters a dataset contains. Of course, previous experiments, preliminary examination of the datasets or even the use of "training" datasets could be used to fix the number of clusters for a given situation (a training dataset is defined as a dataset used to train or build a model). Moreover, the subjective estimation of the number of clusters must not be underestimated, and, even if other methods are used, it might be useful for identifying real patterns and discarding outliers or identifying potentially valuable clusters.

There are many statistics that can be used to estimate the optimal number of clusters in a dataset [18,76]. These statistics are sometimes included in the clustering algorithm, stopping the iterations when a given value is reached, although these rules might have limitations [66]. A useful method for deciding on the

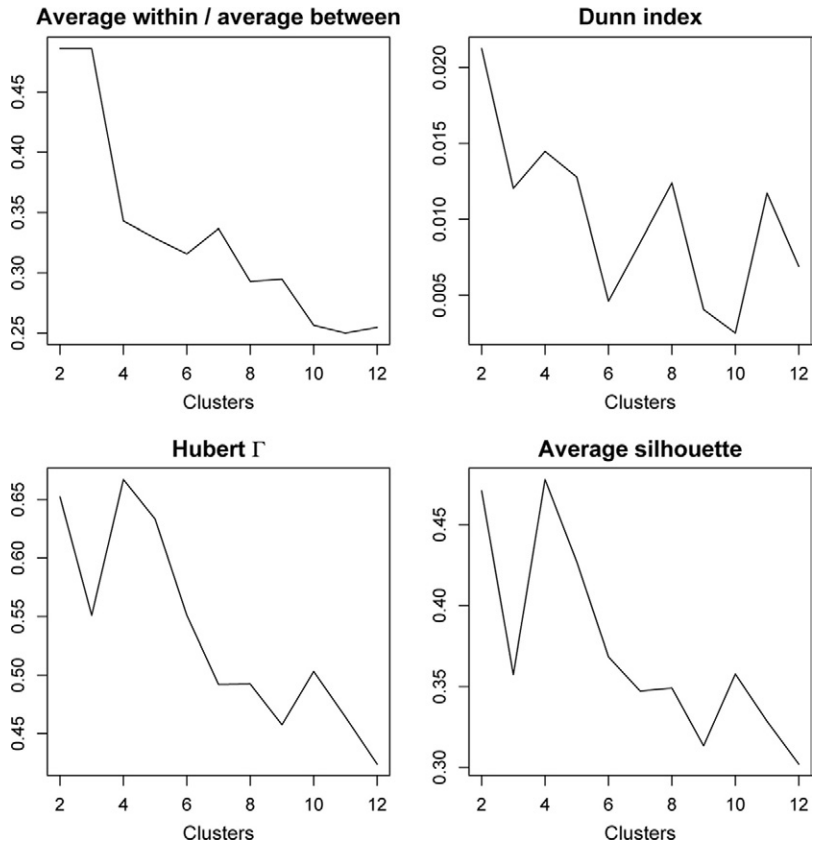

Fig. 2. Example of four indexes for choosing the final number of clusters. A CASA dataset (1000 motile spermatozoa) was processed using the partitional algorithm CLARA [17], with k (final number of clusters) taking values from 2 to 12. For each k value, four validation statistics were calculated: the ratio among the average distances within clusters and between clusters; the Dunn index (ratio among the minimum separation and maximum diameter of the clusters); the Hubert Γ (which assess the compliance between a partitioning and the distance matrix); and the average silhouette width (the silhouette value measures the degree of confidence in the clustering assignment of a particular observation). It is expected that a correct number of clusters would be followed by an increase of the within/between ratio when decreasing the number of clusters. For the other indexes, peaks announce choices that might reflect the adequate cluster structure of the dataset. It is important to realize that more than one choice is possible, and that different indexes might not agree. For instance, the Hubert Γ and the average silhouette indicate that 4 and 10 would be good choices, whereas the Dunn index show 10 clusters as a bad option, suggesting 8 and 11 clusters instead (Dunn index is more sensitive to noise than Hubert Γ or the average silhouette width [76]). Note that 4, 8 and 10 are followed by sharp increases of the within/between ratio.

number of clusters consists in calculating distance-based statistics for decreasing numbers of clusters. Sharp variations in the values of these statistics indicate potentially optimal numbers of clusters. Thus, plotting the number of clusters vs. the value of the statistics is often employed to detect these variations (that look as "peaks" or "valleys", as shown in Fig. 2). Examples of such statistics are the Dunn index and the Hubert's Γ. Model-based methods use a different approach, trying to optimize some criterion functions, such as the Bayesian inference criterion (BIC). Notwithstanding, quoting Yeung and Ruzzo [52], "on purely philosophical grounds, it seems impossible to determine the 'right' number of clusters, or even to define the concept, in the absence of a well-grounded statistical model". These authors stated in 2001 that a well-grounded statistical model was not yet available for gene expression data, and we could say the same for CASA data. This is not to say that subpopulation analysis in CASA data has no real value, but that further research must be done to identify the best methodology to carry out this analysis, and even developing customized ones.

4.4. Validation of the cluster solution

Cluster validation is another important step that should be performed after a clustering analysis. There are many statistical strategies to validate a cluster solution [18,52,76], some of them apt for CASA cluster-

ing. The best option would be to compare our classification to a standard—already classified—dataset, but it might not be possible. The adjusted Rand index or Hubert's Γ can be used to assess the degree of agreement between two classifications [47], being useful not only to compare with a standard, but also to compare different clustering approaches (a sensible strategy, especially when testing the suitability of different algorithms on our dataset). Cophenetic correlations, which measures the similarity between original dissimilarities and dissimilarities estimated from the clustering, can be used to validate the solutions of different algorithms, but only when hierarchical methods are used [66,76].

Once sub-optimal methods are discarded, the stability of the cluster solution can be tested. Clusters should be tested for *quality* (as much compact and separated as possible), *stability* (consistency of the solution even in the presence of disturbances) and, of utmost importance, *biological meaning*. Some statistics for obtaining the optimal number of clusters can be also used for assessing cluster quality, as mentioned in the previous paragraph. Stability can be assessed by subsampling, assuming that random samples from the dataset would produce a consistent clustering solution. Others have proposed the use of bootstrapping or jackknifing (kind of resampling methods) to generate many slightly perturbed datasets from the original data, and testing if the clustering structure remains the same despite of the perturbation [77]. Finally, even if statistically optimal, a cluster solution may lack biological significance. Therefore, researchers must study the results of cluster analysis carefully and, if possible, test the subpopulation pattern against other biological data (sperm physiological parameters or fertility results).

We must point out that in any case it is correct to compare the average values of the clusters (using *p* values) to probe the validity of a clustering solution. Clustering algorithms are best suited to separate cases among distinct groups, and therefore, finding significant *p* values after such a comparison is not a proof of the validity of the classification, but something expected!

5. Working with the subpopulations

After assigning the observations (spermatozoa) to clusters (subpopulations), the latter can be characterized by calculating the respective average values of the CASA parameters (the median and other robust estimates should be preferred). The frequencies of each subpopulation within samples, males and/or treatments can be easily calculated, thus obtaining different subpopulation patterns. These population frequencies can be used for carrying out further statistical analyses. Differences in the relative proportions of the subpopulations among samples or treatments may relate to changes in sperm physiology. Similarly, regression or other statistical techniques may be used to relate subpopulations to other sperm features (physiological markers, fertility results, etc.) [39,41], and to test their predictive value. In fact, this is the final aim of subpopulation analysis, where its usefulness can be tested and applied to better understand sperm biology.

Some difficulties can arise when comparing subpopulations obtained from different datasets (even within the same experiment), because the subpopulation characteristics could vary. Most of the time it is easy to find analogies among subpopulations found in different analyses ("slow and non-linear spermatozoa", "fast and linear", etc.), which might be enough for most purposes. Nevertheless, some graphical tools can help when comparing different subpopulation sets. For instance, Chernoff faces can be very effective for finding relationships among objects defined by many numeric parameters (such as CASA clusters), because of the ability of the human brain to deal with face-like entities [24,33].

Finally, it is expected that subpopulation studies advance beyond the unsupervised classification methods described previously. Once the subpopulation patterns are characterized, it would be desirable to use validated datasets as a guide to classify new datasets, instead of repeating the clustering process. This has been attempted in several studies. For instance, Holt [26] used discriminant analysis (a supervised classification system) to assign cluster memberships to unclustered datasets, using an initial dataset that had been classified using cluster analysis. Data mining and machine learning are disciplines that can contribute enormously to this purpose. Machine learning is aimed at design computer programs that solve a task not based on predefined rules provided by the user, but using relations that they 'learned' from the information, data or feedback that they receive [78,79]. A dataset (the clustered one) is used as a *training sample*. Using that dataset, the system builds a prediction model (*learner*), enabling prediction of the outcome for new observations, the *test set*. The application of these statistical tests on validated clustered datasets would allow the development of new applications based in CASA-defined subpopulations.

6. Future perspectives and conclusions

Use of sperm subpopulation analysis is not yet widespread. Even though the availability of CASA for motility assessment has increased and its use has been reported in almost 500 articles in the last 10 years, only around 30 of them reported using cluster analysis to detect sperm subpopulations. Indeed, while other areas have received dedicated attention from the Statistical Sciences [78,80], sperm motility analysis by multivariate methods has not spread beyond reproductive journals yet. There are promising results, but many aspects must be studied in detail (e.g., the internal characteristics of CASA datasets, cluster shapes, algorithm optimization, variable selection methods, etc.). Moreover, it is necessary to perform comparative studies among several statistical approaches, using both real and simulated datasets. We should be able to define the typical clustering structures of sperm from different species, at physiological or non-physiological conditions, while selecting a set of robust and validated statistical protocols to process these data. As mentioned previously, it may be desirable that, once studies on unsupervised classification are able to characterize clusters with a defined biological meaning, we apply machine learning techniques [78], to develop automated supervised classification of sperm samples. Given the increasing capability of computer hardware, CASA software could be upgraded to yield reliable information about sperm subpopulations in real time.

To achieve these goals, we must also consider three factors. First, CASA systems, both hardware (specialized optics, high definition cameras, fast computers) and software (fast and bug-free, improved track-resolving algorithms) must continue improving. It is necessary that developers follow the minimal requirements, improvements and standards demanded by experts [10–14]. CASA data are greatly affected by variations among software algorithms (to obtain VAP, ALH, etc.) and acquisition settings. Therefore, researchers should adhere to the aforementioned recommendations as closely as possible, detailing in their reports the settings of their systems with great detail, and CASA developers should provide as much information as possible about their products.

Second, open source solutions should be developed and be available, together with proprietary options. Open source software has been successfully applied in many biological fields, helping to advance the informatics tools and to standardize methods and formats [81–83]. Advances in open source bioimaging software has led to promising efforts to develop an open source CASA [15], and there are many open source clustering options [81,84,85]. Open source software provides the source code of the applications, allowing one to examine directly the algorithms that process the data, facilitating collaborative development and feedback. The presence of open source projects usually helps enforce standards (instead having to deal with proprietary data formats, which need to be processed only with a specific CASA software), and promote sharing information and standardization on image processing and statistical algorithms.

Lastly, sperm subpopulations must be thoroughly characterized, identifying those patterns that have a consistent biological meaning. As indicated in the Introduction, several authors have made attempts to relate specific subpopulations to the freezability or fertility of sperm samples. It is necessary to confirm such relationships, and to perform molecular analyses to study if spermatozoa in these subpopulations share specific physiological traits. Moreover, only few studies have considered the changes of motility subpopulations during capacitation [23,26,56]. Since the fertility of a semen sample requires the presence of a subpopulation responsive to the oviductal environment [64], the changes in the subpopulation pattern must be linked to these physiological changes and, ultimately, to the fertility of the semen samples. To achieve this task, we must take into account that artificial reproductive techniques may modify both motility patterns and the relationship among subpopulations and fertility outcomes. Future studies should characterize sperm samples and consider different approaches regarding species, source of the samples (epididymal, semen in seminal plasma, spermatozoa diluted/washed/selected), storage (cooled, cryopreserved), and application (vaginal/transcervical or intrauterine-laparoscopic insemination, *in vitro* fertilization, etc.). It is possible that subpopulation patterns have different meanings in each of these contexts, and research should focus in deciphering these meanings.

Acknowledgments

F. Martínez-Pastor was supported by the Juan de la Cierva program and by the Ramón y Cajal program (Ministry of Science and Innovation, Spain).

References

[1] Tash J, Bracho G. Identification of phosphoproteins coupled to initiation of motility in live epididymal mouse sperm. Biochem Biophys Res Commun 1998;251:557–63.

[2] Hamamah S, Gatti J. Role of the ionic environment and internal pH on sperm activity. Hum Reprod 1998;13 Suppl 4:20–30.
[3] Chamberland A, Fournier V, Tardif S, Sirard M, Sullivan R, Bailey J. The effect of heparin on motility parameters and protein phosphorylation during bovine sperm capacitation. Theriogenology 2001;55:823–35.
[4] Boyers SP, Davis R, Katz D. Automated semen analysis. Curr Probl Obstet Gynecol Fertil 1989;12:172–200.
[5] Katz DF, Erickson RP, Nathanson M. Beat frequency is bimodally distributed in spermatozoa from t/t12 mice. J Exp Zool 1979;210:529–35.
[6] Neill J, Olds-Clarke P. A computer-assisted assay for mouse sperm hyperactivation demonstrates that bicarbonate but not bovine serum albumin is required. Gamete Res 1987;18: 121–40.
[7] Katz DF, Davis RO. Automatic analysis of human sperm motion. J Androl 1987;8:170–81.
[8] Chantler E, Abraham-Peskir J, Roberts C. Consistent presence of two normally distributed sperm subpopulations within normozoospermic human semen: a kinematic study. Int J Androl 2004;27:350–9.
[9] Slott VL, Suarez JD, Perreault SD. Rat sperm motility analysis: methodologic considerations. Reprod Toxicol 1991;5:449–58.
[10] Davis RO, Katz DF. Standardization and comparability of CASA instruments. J Androl 1992;13:81–6.
[11] Mortimer D, Aitken RJ, Mortimer ST, Pacey AA. Workshop report: clinical CASA–the quest for consensus. Reprod Fertil Dev 1995;7:951–9.
[12] ESHRE Andrology Special Interest Group. Guidelines on the application of CASA technology in the analysis of spermatozoa. Hum Reprod 1998;13:142–5.
[13] Mortimer ST. CASA–practical aspects. J Androl 2000;21: 515–24.
[14] Amann RP, Katz DF. Reflections on CASA after 25 years. J Androl 2004;25:317–25.
[15] Wilson-Leedy JG, Ingermann RL. Development of a novel CASA system based on open source software for characterization of zebrafish sperm motility parameters. Theriogenology 2007;67:661–72.
[16] Katkov I, Lulat AG. Do conventional CASA-parameters reflect recovery of kinematics after freezing? CASA paradox in the analysis of recovery of spermatozoa after cryopreservation. Cryo Letters 2000;21:141–8.
[17] Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley-Interscience, Hoboken NJ. 1990.
[18] Xu R, Wunsch D 2nd. Survey of clustering algorithms. IEEE Trans Neural Netw 2005;16:645–78.
[19] Everitt BS, Landau S, Leese M. Cluster Analysis. Wiley, Hoboken, NJ, 4th ed. 2009.
[20] Aggarwal RAK, Ahlawat SPS, Kumar Y, Panwar PS, Singh K, Bhargava M. Biometry of frozen-thawed sperm from eight breeds of Indian buffaloes (*Bubalus bubalis*). Theriogenology 2007;68:682–6.
[21] Álvarez M, Garcìa-Macìas V, Martìnez-Pastor F, Martìnez F, Borragán S, Mata M, Garde J, Anel L, De Paz P. Effects of cryopreservation on head morphometry and its relation with chromatin status in brown bear (*Ursus arctos*) spermatozoa. Theriogenology 2008;70:1498–506.
[22] Esteso MC, Fernandez-Santos MR, Soler AJ, Montoro V, Martinez-Pastor F, Garde JJ. Identification of sperm-head morphometric subpopulations in Iberian red deer epididymal sperm samples. Reprod Domest Anim 2009;44:206–11.
[23] Davis RO, Overstreet JW, Asch RH, Ord T, Silber SJ. Movement characteristics of human epididymal sperm used for fertilization of human oocytes in vitro. Fertil Steril 1991;56: 1128–35.
[24] Davis R, Siemers R. Derivation and reliability of kinematic measures of sperm motion. Reprod Fertil Dev 1995;7:857–69.
[25] Davis R, Drobnis E, Overstreet J. Application of multivariate cluster, discriminate function, and stepwise regression analyses to variable selection and predictive modeling of sperm cryosurvival. Fertil Steril 1995;63:1051–7.
[26] Holt W. Can we predict fertility rates? Making sense of sperm motility. Reprod Domest Anim 1996;31:17–24.
[27] Abaigar T, Holt W, Harrison R, del Barrio G. Sperm subpopulations in boar (*Sus scrofa*) and gazelle (*Gazella dama mhorr*) semen as revealed by pattern analysis of computer-assisted motility assessments. Biol Reprod 1999;60:32–41.
[28] Nunez-Martinez I, Moran J, Pena F. A three-step statistical procedure to identify sperm kinematic subpopulations in canine ejaculates: changes after cryopreservation. Reprod Domest Anim 2006;41:408–15.
[29] Martinez-Pastor F, Garcia-Macias V, Alvarez M, Herraez P, Anel L, de Paz P. Sperm subpopulations in Iberian red deer epididymal sperm and their changes through the cryopreservation process. Biol Reprod 2005;72:316–27.
[30] Quintero-Moreno A, Miro J, Rigau T, Rodriguez-Gil JE. Identification of sperm subpopulations with specific motility characteristics in stallion ejaculates. Theriogenology 2003; 59:1973–90.
[31] Martinez-Pastor F, Guerra C, Kaabi M, Diaz AR, Anel E, Herraez P, de Paz P, Anel L. Decay of sperm obtained from epididymes of wild ruminants depending on postmortem time. Theriogenology 2005;63:24–40.
[32] Martinez-Pastor F, Diaz-Corujo A, Anel E, Herraez P, Anel L, de Paz P. Post mortem time and season alter subpopulation characteristics of Iberian red deer epididymal sperm. Theriogenology 2005;64:958–74.
[33] Martinez-Pastor F, Cabrita E, Soares F, Anel L, Dinis MT. Multivariate cluster analysis to study motility activation of *Solea senegalensis* spermatozoa: a model for marine teleosts. Reproduction 2008;135:449–59.
[34] Toth GP, Stober JA, Read EJ, Zenick H, Smith MK. The automated analysis of rat sperm motility following subchronic epichlorohydrin administration: methodologic and statistical considerations. J Androl 1989;10:401–15.
[35] Filzmoser P, Garrett RG, Reimann C. Multivariate outlier detection in exploration geochemistry. Comp Geosci 2005;31: 579–87.
[36] Fraley C, Raftery A. Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 2002;97:611–32.
[37] Filzmoser P, Hron K, Reimann C. Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. Sci Total Environ 2009;407:6100–8.
[38] Steinley D, Brusco M. Selection of variables in cluster analysis: An empirical comparison of eight procedures. Psychometrika 2008;73:125–44.
[39] Quintero-Moreno A, Rigau T, Rodriguez-Gil JE. Regression analyses and motile sperm subpopulation structure study as improving tools in boar semen quality analysis. Theriogenology 2004;61:673–90.

[40] Miró J, Lobo V, Quintero-Moreno A, Medrano A, Peña A, Rigau T. Sperm motility patterns and metabolism in Catalonian donkey semen. Theriogenology 2005;63:1706–16.
[41] Quintero-Moreno A, Rigau T, Rodriguez-Gil J. Multivariate cluster analysis regression procedures as tools to identify motile sperm subpopulations in rabbit semen and to predict semen fertility and litter size. Reprod Domest Anim 2007;42:312–9.
[42] Flores E, Fernández-Novell JM, Peña A, Rodrìguez-Gil JE. The degree of resistance to freezing-thawing is related to specific changes in the structures of motile sperm subpopulations and mitochondrial activity in boar spermatozoa. Theriogenology 2009;72:784–97.
[43] Miró J, Taberner E, Rivera M, Peña A, Medrano A, Rigau T, Peñalba A. Effects of dilution and centrifugation on the survival of spermatozoa and the structure of motile sperm cell subpopulations in refrigerated Catalonian donkey semen. Theriogenology 2009;72:1017–22.
[44] Rivera MM, Quintero-Moreno A, Barrera X, Palomo MJ, Rigau T, Rodrìguez-Gil JE. Natural Mediterranean photoperiod does not affect the main parameters of boar-semen quality analysis. Theriogenology 2005;64:934–46.
[45] Rivera MM, Quintero-Moreno A, Barrera X, Rigau T, Rodrìguez-Gil JE. Effects of constant, 9 and 16-h light cycles on sperm quality, semen storage ability and motile sperm subpopulations structure of boar semen. Reprod Domest Anim 2006;41:386–93.
[46] Rodrìguez-Gil JE, Silvers G, Flores E, Jesús Palomo M, Ramìrez A, Montserrat Rivera M, Castro M, Brito M, Bücher D, Correa J, Concha II. Expression of the GM-CSF receptor in ovine spermatozoa: GM-CSF effect on sperm viability and motility of sperm subpopulations after the freezing-thawing process. Theriogenology 2007;67:1359–70.
[47] Brusco M, Cradit J. A variable-selection heuristic for k-means clustering. Psychometrika 2001;66:249–70.
[48] Huang JZ, Ng MK, Rong H, Li Z. Automated variable weighting in k-means type clustering. IEEE Trans Pattern Anal Mach Intell 2005;27:657–68.
[49] Rigau T, Farre M, Ballester J, Mogas T, Pena A, Rodriguez-Gil JE. Effects of glucose and fructose on motility patterns of dog spermatozoa from fresh ejaculates. Theriogenology 2001;56:801–15.
[50] Martinez-Pastor F, Garcia-Macias V, Alvarez M, Chamorro C, Herraez P, de Paz P, Anel L. Comparison of two methods for obtaining spermatozoa from the cauda epididymis of Iberian red deer. Theriogenology 2006;65:471–85.
[51] Dorado J, Molina I, Muñoz-Serrano A, Hidalgo M. Identification of sperm subpopulations with defined motility characteristics in ejaculates from Florida goats. Theriogenology 2010;74:795–804.
[52] Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. Bioinformatics 2001;17:763–74.
[53] King JR, Jackson DA. Variable selection in large environmental data sets using principal components analysis. Environmetrics 1999;10:67–77.
[54] Al-Kandari NM, Jolliffe IT. Variable selection and interpretation in correlation principal components. Environmetrics 2005;16:659.
[55] Ramio L, Rivera MM, Ramirez A, Concha II, Pena A, Rigau T, Rodriguez-Gil JE. Dynamics of motile-sperm subpopulation structure in boar ejaculates subjected to "in vitro" capacitation and further "in vitro" acrosome reaction. Theriogenology 2008;69:501–12.
[56] Muiño R, Tamargo C, Hidalgo CO, Peña AI. Identification of sperm subpopulations with defined motility characteristics in ejaculates from Holstein bulls: effects of cryopreservation and between-bull variation. Anim Reprod Sci 2008a;109:27–39.
[57] Flores E, Taberner E, Rivera MM, Peña A, Rigau T, Miró J, Rodríguez-Gil JE. Effects of freezing/thawing on motile sperm subpopulations of boar and donkey ejaculates. Theriogenology 2008;70:936–45.
[58] Muiño R, Rivera MM, Rigau T, Rodriguez-Gil JE, Peña AI. Effect of different thawing rates on post-thaw sperm viability, kinematic parameters and motile sperm subpopulations structure of bull semen. Anim Reprod Sci 2008b;109:50–64.
[59] Corral-Baqués MI, Rivera MM, Rigau T, Rodríguez-Gil JE, Rigau J. The effect of low-level laser irradiation on dog spermatozoa motility is dependent on laser output power. Lasers Med Sci 2009;24:703–13.
[60] Muiño R, Peña AI, Rodríguez A, Tamargo C, Hidalgo CO. Effects of cryopreservation on the motile sperm subpopulations in semen from Asturiana de los Valles bulls. Theriogenology 2009;72:860–8.
[61] van Ooyen A. Theoretical aspects of pattern analysis. In: L Dijkshoorn, KJ Towner, M Struelens, editors, New Approaches for the Generation and Analysis of Microbial Typing Data, Elsevier, Amsterdam. p. 31–45.
[62] Holt W, Harrison R. Bicarbonate stimulation of boar sperm motility via a protein kinase A-dependent pathway: between-cell and between-ejaculate differences are not due to deficiencies in protein kinase A activation. J Androl 2002;23:557–65.
[63] Satake N, Elliott RMA, Watson PF, Holt WV. Sperm selection and competition in pigs may be mediated by the differential motility activation and suppression of sperm subpopulations within the oviduct. J Exp Biol 2006;209:1560–72.
[64] Steinbach M, Ertoz L, Kumar V. The challenges of clustering high dimensional data. In: LT Wille, editor. New directions in statistical physics: econophysics, bioinformatics, and pattern recognition, Springer, Berlin. p. 273–310.
[65] Leonard ST, Droege M. The uses and benefits of cluster analysis in pharmacy research. Res Social Adm Pharm 2008;4:1–11.
[66] Zhang T, Ramakrishnan R, Livny M. Birch: An efficient data clustering method for very large databases. In: HV Jagadish, IS Mumick, editors, Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4−6, 1996. ACM Press, 1996, p. 103–14.
[67] Martinez IN, Moran JM, Pena FJ. Two-step cluster procedure after principal component analysis identifies sperm subpopulations in canine ejaculates and its relation to cryoresistance. J Androl 2006;27:596–603.
[68] Ortega-Ferrusola C, Macías García B, Suárez Rama V, Gallardo-Bolaños JM, González-Fernández L, Tapia JA, Rodrìguez-Martinez H, Peña FJ. Identification of sperm subpopulations in stallion ejaculates: changes after cryopreservation and comparison with traditional statistics. Reprod Domest Anim 2009;44:419–23.
[69] Abaigar T, Cano M, Pickard AR, Holt WV. Use of computer-assisted sperm motility assessment and multivariate pattern analysis to characterize ejaculate quality in Mohor gazelles (*Gazella dama mhorr*): effects of body weight, electroejaculation technique and short-term semen storage. Reproduction 2001;122:265–73.
[70] Cremades T, Roca J, Rodriguez-Martinez H, Abaigar T, Vazquez JM, Martinez EA. Kinematic changes during the cryopreservation of boar spermatozoa. J Androl 2005;26:610–8.

[71] Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput J 1998;41:578–88.
[72] Fraley C, Raftery AE. Enhanced model-based clustering, density estimation, and discriminant analysis software: Mclust. J Classif 2003;20:263–86.
[73] Fraley C, Raftery A, Wehrens R. Incremental model-based clustering for large datasets with small clusters. J Comput Graph Stat 2008;14:529–46.
[74] Domínguez-Rebolledo AE, Fernández-Santos MR, García-Alvarez O, Maroto-Morales A, Garde JJ, Martínez-Pastor F. Washing increases the susceptibility to exogenous oxidative stress in red deer spermatozoa. Theriogenology 2009;72:1073–84.
[75] Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. Bioinformatics 2005;21:3201–12.
[76] Kerr MK, Churchill GA. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci U S A 2001;98:8961–5.
[77] Tarca AL, Carey VJ, Chen Xw, Romero R, Drãghici S. Machine learning and its applications to biology. PLoS Comput Biol 2007;3:e116.
[78] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. Springer Series in Statistics. Springer-Verlag, Berlin, second ed. 2009.
[79] Kuo WP, Kim EY, Trimarchi J, Jenssen TK, Vinterbo SA, Ohno-Machado L. A primer on gene expression and microarrays for machine learning researchers. J Biomed Inform 2004;37:293–303.
[80] Reimers M, Carey VJ. Bioconductor: an open source framework for bioinformatics and computational biology. Methods Enzymol 2006;411:119–34.
[81] Stajich JE, Lapp H. Open source tools and toolkits for bioinformatics: significance, and where are we? Brief Bioinform 2006;7:287–96.
[82] Swedlow JR, Eliceiri KW. Open source bioimage informatics for cell biology. Trends Cell Biol 2009;19:656–60.
[83] de Hoon MJL, Imoto S, Nolan J, Miyano S. Open source clustering software. Bioinformatics 2004;20:1453–4.
[84] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2008.
[85] Beirão J, Soares F, Herráez MP, Dinis MT, Cabrita E. Sperm quality evaluation in *Solea senegalensis* during the reproductive season at cellular level. Theriogenology 2009;72:1251–61.